引用格式：滕妍, 王国豫, 王迎春. 通用模型的伦理与治理：挑战及对策. 中国科学院院刊, 2022, 37(9): 1290-1299.
Teng Y, Wang G Y, Wang Y C. Ethics and governance of general models: Challenges and countermeasures. Bulletin of Chinese Academy of Sciences, 2022, 37(9): 1290-1299. (in Chinese)

通用模型的伦理与治理：挑战及对策

滕　妍[1]　王国豫[2,3*]　王迎春[1]

1 上海人工智能实验室 治理研究中心 上海 200232
2 复旦大学 哲学学院 上海 200433
3 复旦大学生命医学伦理研究中心 上海 200433

摘要　通用模型是近年来人工智能发展的重要方向之一。随着模型研发应用的增多，模型的社会和伦理影响受到广泛关注。文章从通用模型的特性出发，分析了模型在算法、数据和算力3个层面潜在的伦理挑战，包括不确定性、真实性、可靠性，偏见、毒性、公平、隐私及环境问题。进一步从技术哲学的视角分析了数据驱动的模型在人与世界关系中的中介性作用及所产生的“镜像”效应问题和透明性问题，提出了人与世界关系的新形态是以模型（数据）为中介的，即“人-模型（数据）-世界”关系。最后，从治理技术和治理机制两方面反思了当前的应对措施及局限性。建议建立开放式、全流程、价值嵌入的伦理规约机制，保障通用模型在合规、合伦理的框架下发展。

关键词　人工智能，通用模型，风险，治理，科技伦理

DOI　10.16418/j.issn.1000-3045.20220505003

1 人工智能步入“通用模型时代”

模型，是现实世界中事件、系统、行为或者自然现象的抽象数学表征[1]。在人工智能领域，通用模型，又称为预训练模型、基础模型或大模型，是指经过大规模数据训练，包含海量模型参数，并可适用于广泛下游任务的模型[2]。自21世纪早期深度学习技术兴起，人工智能的研发模式主要是开发专用模型，即针对特定任务、场景，依赖大量人力的“作坊式”。然而，由于专用模型泛化能力差且严重依赖专用数据集，导致开发成本高[3]。近几年，人们开始聚焦通用模型，即通过预先训练一个超大规模的模型，再用少量微调数据将模型迁移到具体场景中，一次性解决下游碎片化任务。

在目前的通用模型中，自然语言处理（NLP）和计算机视觉（CV）是发展最快的两个领域。NLP指

*通信作者
资助项目：上海人工智能实验室，上海市科学技术委员会项目（21511104503、21511104502），国家自然科学基金（L2124040）
修改稿收到日期：2022年8月31日

的是建立在计算机编程基础上的人与计算机之间的自然语言通讯，即让人工智能“说人话”[4]，代表模型包括Bert（Google）、GPT-2/3（OpenAI）、GLaM（Google）、“盘古”（华为）、“悟道”（智源、清华）等。CV指的是利用传感器、摄像头、计算机等设备模仿生物视觉能力，即让人工智能“看世界”，代表包括Swin Transformer（Microsoft）、CLIP（OpenAI）、“书生”（上海人工智能实验室）等。

从技术的角度来看，通用模型所基于的深度神经网络和自监督学习并不新颖，其近年来的迅速发展主要归功于不断扩大的模型参数量、训练数据及算力的发展。比如，OpenAI在2018年发布的语言模型GPT-1的参数量为1.1亿，而其在2020年发布的GPT-3模型的参数量已高达1 750亿，而Google的Switch Transformer和北京智源研究院的“悟道”2.0的参数已经达到万亿级别[3]。模型参数、训练数据和算力的扩增激发了模型的潜能，提高了模型的泛化性和适用性，提升了开发效率。*在通用性方面，*据华为数据显示，其“盘古”通用模型可以适配电力行业的100多种问题，迭代效率提升了10倍左右①。*在精度方面，*Clark等[5]的实验显示，在没有训练的情况下，人类评价者已经很难区分测试文本是GPT-3生成的还是人类创作的。*在数据效率方面，*上海人工智能实验室的“书生”模型仅需2个训练样本，就能在花卉种类识别的任务上达到99.7%的精确度[6]。

通用性能和模型质量的提升降低了人工智能的研发和使用门槛，促进了人工智能的规模化落地，未来可能带来从技术、规则、制度到产业的整个人工智能领域的范式转变。据GPT-3官网显示，截至2022年2月，已经有超过320个APP采用了GPT-3模型，其中包括人们熟知的WhatsApp、Duolingo、Excel等产品②。人工智能已然步入“通用模型”时代。

2 通用模型的特性及伦理挑战

2.1 通用模型的特性

2.1.1 核心特性

通用模型的核心特性可以归纳为涌现性（emergence）和同质性（homogenization）[2]。

（1）*涌现性。*涌现是复杂系统的特征，通常指的是“在复杂系统的自组织过程中，出现新颖而连贯的结构、模式和属性”[7]。涌现具有2个明显特性：① 单体之间相互作用产生的协同效应，这使群体作为一个整体可以产生个体简单相加无法达到的效果。② 在临界状态下质的改变，这种质变的例子包括自然界中的雪崩、冰化成水，又或者经济领域的股市价格波动等[8]。*在通用模型的语境下，涌现性指的是模型的功能不是具体建构的，而是在模型的大规模训练过程中自主涌现出来的。*通用模型的功能涌现性根植于机器学习的特性。机器学习领域的先锋——Samuel[9]指出，机器学习算法的本质是让计算机自动学习训练数据的内在规律，并用这些规律对未知数据进行预测，这种基于训练数据产生模型的算法具有涌现性的特征。通用模型在大规模算力和数据的加持下，不仅可以自动指定任务的执行方式、提取数据的内在特征，还可以涌现出一些预料之外的功能[2]。Steinhardt[10]将通用模型的涌现性进一步解释为由于参数数量增加而引起的行为上的质变，这种质变可以引起系统的快速改变，在产生新功能的同时也可能产生新的风险。

（2）*同质性。*同质性，与异质性（heterogenization）相反，指的是构成物体或人的单元在质量、结构、

① 谢凌曦.《盘古预训练大模型》和我们对大模型的思考.(2021-07-07)[2022-06-29]. http://valser.org/webinar/slide/slides/20210707/0707-VALSE-%E5%A4%A7%E6%A8%A1%E5%9E%8B-%E8%B0%A2%E5%87%8C%E6%9B%A6.pdf.

② 数据来源：GPT-3应用官网（https://gpt3demo.com/）。

状态等特征上的单一或相似性[③]。在通用模型语境下，同质性指的是模型构建方法的相似性。主要表现在 2 个方面：① 横向的同质化，指的是类似的网络架构（如 Transformer）已经被广泛用于诸多领域的模型开发中。如文本、图片、视频、蛋白质结构预测等，使得模型的能力可以惠及诸多领域。② 纵向的同质化，指的是类似或同一模型被广泛用于解决下游任务。通用模型领域模型建构的同质化主要源于模型强泛化性所带来的开发、优化、应用等方面效率的提升。此外，通用模型作为一种中间模型，其问题也会被下游模型所继承，带来“一荣俱荣，一损俱损”的效果。

2.1.2 衍生特性

通用模型还有两个衍生特性：大规模数据训练和任务泛化。① 大规模数据训练不仅是模型功能涌现性的重要前提，也对提高模型整体适用性、准确度和效率发挥了积极作用。② 通用模型任务泛化的特性，指的是模型可以作为具有“通才”能力的中间模型，被使用者迁移处理各种不同的、甚至未知的任务，而类似网络架构在多种模态模型（文字、图片、视频等）中的运用更是拓宽了模型的下游应用领域。在这个意义上，通用模型可以被看作是一种促能技术（enabling technology）。

2.2 通用模型的伦理挑战

通用模型的上述特性推进了人工智能技术的发展和规模化落地，但也带来一系列伦理挑战，包括涌现性伴随的功能不确定、鲁棒性欠缺所导致的可靠性问题与认知风险；大规模数据训练面临的隐私安全、偏见、毒性、公平性问题；同质性和任务泛化对上述问题在横向和纵向上的放大和加深；使用海量算力对能源的消耗和环境的影响。总的来说，通用模型面临的伦理挑战主要表现在 3 个层面。

2.2.1 算法层面：不确定性与真实性、可靠性问题

从认识论角度看，通用模型的涌现性直接呈现出的是鲁棒性欠缺，导致算法的不确定性和不可解释性，其结果也影响到算法的真实性和可靠性，而同质性则放大了这些问题在横向和纵向上的影响。

Hendrycks 等[11]将机器学习的鲁棒性解释为系统面对黑天鹅事件和对抗性威胁时表现出的能力。以交通灯为例，遵守信号灯行驶在绝大多数场景下是安全的，但是也有极少数特殊情况。比如，当十字路口横向来车闯红灯时，纵向行驶的车辆如若还按绿灯行驶就会造成车祸。考虑到通用模型的通用性质，现阶段大部分模型在制定安全性指标时会将一套复杂的人工智能安全性原则汇总到一个概括性的综合指标中，而不对其进行更细粒度的区分；但在具体应用场景中，往往需要考虑更细粒度的安全问题，如信号灯故障及其他突发、极端情况。

不同于黑天鹅事件，对抗性威胁通常是针对算法或模型漏洞和弱点的故意性攻击，操作方式通常是刻意在样本中加入人类无法察觉的干扰[12]。2015 年，谷歌的研究人员发现，对抗样本可以轻易地欺骗 GoogLeNet 模型。研究人员制作了两张熊猫图片，并在其中一幅的像素层面加入细微干扰作为对抗样本，对人类肉眼来说，这两张图几乎是一模一样，而机器学习模型却以高置信度将对抗样本误读为长臂猿[13]。类似的情况还有，OpenAI 的研究人员发现其实验室最新的通用视觉模型可以轻易被欺骗。例如，如果在一颗青苹果上贴上“ipod”标签，那该模型有 99.7% 的概率将图中物体识别成 ipod，而不是青苹果[14]。

产生对抗威胁的直接原因是模型与人类处理信息方式的不同，人类是基于各种可理解的特征来识别图片中的内容，如熊猫的黑眼圈、黑耳朵等。而模型的识别方法是基于对大规模数据样本的拟合学习，得出

③ Cambridge Dictionary. [2022-06-29]. https://dictionary.cambridge.org/dictionary/english/homogeneity.

一个可以将所有图片与标签相关联的参数框架，然后再将这种关联泛化到其他任务中。但是这种关联本质上源于模型的涌现性，具有不确定性，有时是不可靠的。而由于通用模型作为“中介”和“促能”技术的重要意义，在下游的应用中起到了“母体”的作用，模型的细小误差可能对下游大规模应用带来影响。

究其根本，产生对抗性威胁的深层原因是模型无法从事实上区分信息的真实性，从而产生虚假的、误导性的或低质量的结果。并且，深度神经网络架构可能使每次推理结果都不同且可解释性低，这不仅对了解模型能做什么、为什么这么做及如何做带来困难，更是为模型的审核与合规带来了一系列挑战[2]。“大数据+海量参数+大算力”提升了模型的总体表现，但是似乎并没有让模型做到像人一样的理解。这在类似下棋等娱乐场景并不会对人类生活产生负面影响。但是，对于诸如医疗诊断、刑侦司法、自动驾驶等重要领域，模型功能的涌现性、应用的同质化和可解释问题可能带来潜在风险。

2.2.2 数据层面：隐私安全、偏见、毒性与公平性问题

现阶段，通用模型在数据层面的主要伦理风险分 2 类：① 训练样本中含有的偏见和毒性（toxicity）问题；② 数据所涉及的隐私安全问题。

（1）通用模型训练所用的数据量大且多为无标注数据，样本中易存在偏见和毒性问题，可能对个体或群体带来歧视和不公，且通用模型的同质化特征可能会放大或深化这些问题。偏见可以理解为一种态度或行为上的倾向，在现实生活中常表现为对特定人群的区别对待，可能造成直接或间接的歧视和不公。美国国家标准与技术研究院将人工智能偏见分为 3 类：① 系统偏见，指由历史、社会造成的对某些群体的偏袒或贬低；② 统计偏见，指由于样本的代表性缺陷而导致的统计上的偏差；③ 人类偏见，指人类决策者思维的系统性错误，这种偏见通常是隐式的、与过往经验相联系[15]。毒性或有害言论与偏见类似，但特指粗鲁、不尊重或不合理的言论，如侮辱、仇恨、脏话、人身攻击、性骚扰等④。从来源上说，偏见和有害语言都是复杂的历史产物，在人类社会中难以彻底消除；但如不加以注意，人工智能就可能会拓宽、深化、合并甚至标准化这些问题，或者带来新的问题。在通用模型的语境下，语料库和模型输出的毒性、偏见程度存在明显相关性。当样本中某些群体被缺乏、不实或过度代表时，模型结果就可能使部分人群——特别是边缘群体遭受不公正的对待，继承或加深社会刻板印象。例如，OpenAI 在用 ZS CLIP 模型鉴别 10 000 张 FairFace 数据库的图片时，结果显示，通用模型对不同人群（如种族、性别、年龄等）的识别是不同的，如更容易将男性和 20 岁以下人群视为罪犯，更容易将女性识别为保姆[16]。

（2）隐私安全也是通用模型在数据层面的主要伦理挑战之一，包括数据泄露、功能僭变（function creep）、伪造、诈骗等具体问题。隐私是数字化世界的主要伦理问题之一[17]。通用模型研发和应用涉及的隐私风险主要来源于原始训练数据中包含的个人信息和模型强大的推理能力[18]。通用模型训练多采用网络公开数据（如网页、论坛、论文、电子书等），其中可能包含姓名、家庭住址、电话号码等个人信息，甚至包含生物识别、宗教、医疗、行踪轨迹等敏感信息，这些信息的泄露不仅违法，且直接侵害到个人的尊严、人身和财产安全。通用模型的“一损俱损”效应也意味着，如果模型非法或不当使用了隐私数据，下游所有应用都会承担这种隐私风险[2]。恶意使用者还可能将泄露信息进一步用于伪造、身份冒用、诈骗等不法目的，造成功能僭变（即泄露或超出原目的使用）。如近几年兴起的深度伪造技术，据 Sensity 的调

④ Jigsaw. About the API. [2022-06-29]. https://developers.perspectiveapi.com/s/about-the-api.

查显示，在公开发布的深伪作品中，96% 都是虚假成人内容，即将网络图片合成到情色视频或图片上，而受害者绝大多数都是女性[19]。

2.2.3 算力层面：可持续发展与全球正义

*在全球“双碳”背景下，模型在算力层面所产生的能源和环境影响尤为值得重视。*为应对气候危机，2015 年，178 个缔约方签署了《巴黎协定》，将控制气温上升作为长期目标，我国也于 2020 年明确了“双碳”目标。近年来，“军备竞赛式”的通用模型研发所造成的环境问题已经备受瞩目[3]。Strubell 等[20]梳理了几种常用的深度学习语言模型在训练环节的碳足迹和计算花费。结果显示，训练这些模型产生的二氧化碳当量（CO_2e）从 0.012—284 吨不等。如在不做超参数调整的情况下，训练一个基于 Bert 的模型所造成的碳排放大致相当于一次环美飞行所造成的排量；而训练神经架构搜索模型所需的计算费用高达约 90 万—300 万美元不等。除了训练环节，模型在应用过程中的环境问题也不容忽视。图灵奖得主帕特森（David Patterson）等[21]的研究显示，多数公司在运行深度神经网络模型时所消耗的能源和算力要多于模型训练阶段。

*算力背后的能源和环境问题已经引发关于环境后果和全球正义的讨论。*一方面，涉及代内正义，因为享受到通用模型便利的往往是发达国家和人群，而被气候变暖影响最严重的国家往往是落后的发展中国家[22]。另一方面，能源消耗和环境污染也关涉代际正义，即对未来子孙的影响。

3 通用模型的中介性

美国著名计算机科学家格雷（Jim Gray）[23]指出，传统上，科学研究的范式主要是实证的、理论的和计算式的，而随着仿真计算和数据收集工具的发展，大量的数据被收集起来，这也促使新的科学范式——数据密集型科学的产生。

预训练模型本质上也是数据驱动的，数据是模型知识的资料来源，模型的能力来自对大量无标注训练数据中抽象共现模式（co-occurrence patterns）的学习[2]。举例来说，如果“面包与果酱”在训练数据中出现的概率很高，那么在运用该模型进行文本生成的时候，如果用户输入“面包与”，模型会在提示栏中显示“果酱”等其他高频共现词。Bender 等[22]将大语言模型的这种映射行为称为“鹦鹉学舌”。这当中包含两个问题：① 通用模型的“镜像”效应在透过数据表征世界的时候，也会将训练数据中包含的社会问题不加分辨地体现出来。对于通用模型来说，这种能力的核心是在海量无标注数据中学习通用的特征表示，提取出有价值的模式和规律，而不对数据内容加以鉴别，这就意味着，如果训练数据中与某些词汇（如“黑人”）相关联的内容多为负面的或刻板印象，则模型输出的结果就可能直接体现这些问题，并可能作为数字记忆将这些内容固化[2]。② 由于通用模型的映射行为反映的是训练数据的内在关联，而不是全面丰富、动态复杂的生活世界，选择哪些数据集、依据什么样的标准，以及如何界定核心概念都会对结果产生影响，导致模型结果可能存在可靠性问题[24]。以上两个问题反映出数据驱动的模型可能存在的问题，即通用模型作为一种“镜像”可能无法准确反映世界，还会影响人类对世界的认识。

通用模型的这种“镜像”可以借用伊德（Don Ihde）的技术哲学来解释。作为一种技术中介，通用模型可以调节人类认识和解释世界的过程。伊德在探讨技术在人与世界之间关系时提出技术与人的 4 种关系[25]。*在具身关系中，*人类与技术作为一个整体去经验世界（如眼镜），此时技术是人“身体的延伸”，即（人-技术）→世界，有助于放大某些感官意识。具身关系要求技术的知觉透明性，即人类越了解技术，人与技术融合所能产生的效用越大。*在诠释关系中，*人类对世界的认识是经过技术工具表

征转换的（如仪表盘），即人→（技术-世界），人们认识的世界是由技术诠释的，技术是世界的表征。诠释关系要求技术达到一定的发展程度，以精准反映世界。在它异关系中，强调技术的自主性，即当技术具有它者性（如人工智能）、成为认识的客体时，透过技术展现的世界就变成了一种技术的加工物，表述为人→技术-（-世界）。在背景关系中，技术作为一种技术背景，成为日常生活的一部分（如暖气），而人往往是在技术失效时才能意识到它的存在，即人→（技术/世界）。

在通用模型的语境下，伊德的“人-技术-世界”被“人-模型（数据）-世界”所取代。一方面，通用模型在某种程度上可以理解为数据和部分世界的“镜像”，拥有表征相关事物复杂逻辑关系的能力，人类透过模型来认识世界。然而，模型的“镜像”效应也将映射出现实世界中偏见歧视等伦理问题，对公平、自由、人格尊严等伦理价值造成挑战。另一方面，作为“身体的延伸”，人类又可能与模型融合在一起去经验世界，但是融合的实际效用取决于技术的透明性，在通用模型语境下主要表现为可解释性问题，即是否可以从人类的角度理解从输入到输出背后的映射逻辑。这关乎人类是否可以信任模型，包括模型是否可靠，是否能够符合具体的法律法规和伦理规范，是否可以用于科学研究等目的，以及是否可以部署到重要的生活领域，如帮助儿童学习知识，辅助法官判案等。传统上，判定某个人或团体的言论是否可靠的基本前提是他们知道自己在说什么，即理解这些言论的意思和背后映射的价值规范、行为要求和社会影响。但是，深度神经网络的可解释性和透明性较低，更多的是基于数据和参数的规模效应给出的粗略结果。在某种程度上是用相关性代替了因果性，伴随而来的就是上述信任的问题。

综上，通用模型在反映世界和让人理解两方面还存在一些问题，作为人与世界关系的技术中介，通用模型在帮助人类处理大规模下游任务的同时，也带来了一系列的不确定性。通用模型的“镜像”只是对训练数据的镜像，而不完全是现实世界的镜像，且人类还无法完全理解、信任模型的预测逻辑，如若想让模型的中介作用更加合理，人类必须在环路之中发挥更大的作用。

4 治理措施及局限性

通用模型的社会与伦理维度关乎模型未来的技术走向、规则制定、接受度等一系列问题，属于通用模型发展的基础问题。目前，很多企业及高校都已启动针对通用模型社会和伦理影响的研究，并制定系列举措对模型的研发和应用进行治理。

4.1 治理技术：上、下游伦理调节

当前，应对通用模型伦理挑战的技术可以粗略划分为上游手段和下游手段，主要是通过对上、下游训练数据进行伦理调节以降低模型“镜像”效应的潜在负面影响。

上游手段主要对应模型的预训练阶段。最典型的举措就是对训练数据集进行清洗。如，DeepMind 在改进 TransformerXL 模型的表现时，删去了被 Perspective API 注释为有毒的内容[26]。Perspective API 通过志愿者打分的方式来量化线上评论的分数——如果 10 个评分者中有 6 个将评论标记为有毒，就将该评论的毒性标记为 0.6，代表词条文本有 60% 的可能性被认为是有害的⑤，该评估模型目前已被广泛用于处理通用模型的毒性问题。由于毒性言论和偏见可能对特定人群造成直接或间接的不公且可能降低模型精度，改善这些问题对提高模型公平性、可靠性等具有明显益处。

下游手段主要指模型为适应具体任务而进行微调

⑤ Jigsaw. About the API. [2022-06-29]. https://developers.perspectiveapi.com/s/about-the-api.

的阶段。为改善模型在特定领域的表现，多数研究团队倾向于下游处理，即在已经完成预训练的模型的基础上用有指向性的小规模数据集进行微调[27]。目前，OpenAI、Google 等都使用了下游微调的手段来改善模型在特定社会和伦理维度上的表现，以增加模型在价值规范方面的敏感度和“理解力”，提高通用模型作为人与世界中介的伦理合理性[27,28]。

4.2 治理机制：宏观与微观层面并行

通用模型的治理机制主要包括宏观和微观两个层面。

在宏观层面，通用模型治理要符合当前人工智能治理的总体框架。政府、国际组织、企业等都针对人工智能的发展制定了系列治理机制。总的来说，这些机制包括伦理原则（如我国的《新一代人工智能治理原则——发展负责任的人工智能》）、政策战略（如美国的《美国人工智能倡议》）、法律法规（如欧盟的《人工智能法（草案）》）、标准（IEEE 的 P7000 系列）等。通用模型的治理隶属于相关主体发布的人工智能治理框架，同时受到行业组织、技术团体、新闻媒体、公众等多元主体的监督。

在微观层面，有必要针对通用模型的特性制定专门性治理机制。目前通用模型的专题治理以模型研发、应用的头部机构为主导，主要做法是针对模型的发布、应用制定一系列制度措施，包括：① 分阶段开放模型。OpenAI 分 4 个阶段、由小至大开放了 GPT-2，以逐步了解模型特性，减少被滥用的可能性。② 只针对特定对象开放，并与合作者签订法律协议以规定双方的权利和义务。③ 制定模型应用的具体规则。OpenAI 在开放 GPT-3 早期模型时会审查所有使用案例，并及时中止可能产生危害的应用，降低潜在风险[29]。

上述治理机制和技术都在一定程度上降低了通用模型潜在的社会和伦理风险，但是这些措施仍具有一定的局限性。① 现阶段针对通用模型的专题治理机制欠缺系统性，没有与人工智能的总体治理框架紧密结合。如果审查下游应用案例的伦理权力被模型提供者垄断，这种权力的集中不仅会降低伦理审查的效力，也提高了使用者的风险。② 现阶段的治理技术还是相对割裂和零散的，没有将技术与伦理很好的融合。例如，作为应用最广泛的去毒评估模型，Perspective API 的众包评审机制更多是基于统计和描述的方法，缺乏规范性和理论性依据。一方面，有害文本的判断与个人经历、文化背景、内容场景等有很强的关联性，对用户 A 或某个地域的群体不具冒犯性的言论可能会冒犯到用户 B 或另外一个地域的群体。因此，需要在评估之前充分考虑评审规则的合理性、评审员的多样性等问题。另一方面，API 现有的打分方式无法区分有害文本的危害程度。现有的评分方式是基于文本被认为有害的可能性，即虽然一些行为比其他行为更让人感到冒犯，但模型只能区分一个行为的冒犯性与否，而不能体现冒犯行为之间的质量差异[28]。有必要在系统理论规范的指导下，探索更合理的分级机制。

总的来说，现阶段应对措施的局限性与缺乏系统性考虑和缺乏理论性、规范性指导密切关联。通用模型作为训练数据的“镜像”，具有建模复杂问题和映射社会问题的能力，但却存在对现实世界中事实与价值的粗糙处理，导致所谓的“长尾效应”，即从最普遍需求的角度来看，模型整体上表现优异且符合各项指标，但是对于差异化、特殊性、个性化的需求而言，模型仍然存在风险。为了使通用模型输出的结果更符合人类价值观，在模型发展的早期就将伦理理论和规范介入是十分必要的。

5 展望：建立开放式、全流程、价值嵌入的伦理工具

现阶段应对措施的局限性表明，下一阶段的通用模型治理需要将以下 3 个因素纳入考量。

（1）要秉持开放式的态度，充分评估数据和模型的伦理影响和风险点。这样做的前提是建立拥有交

叉学科背景的伦理委员会、透明的评审规则、合理的伦理权力分配机制和第三方审查机制。在此基础上，从多角度评估数据和模型的潜在影响，特别是数据的多样性、文化背景和模型应用的时间和空间属性，充分评估模型在特定场景下、对特定人群可能产生的影响，降低潜在的负面效应。

（2）从全流程的视角出发，将伦理考量纳入通用模型及其生态系统建设进程之中。一方面，要通过教育、培训等手段加强科研人员和使用人员的伦理素养，让相关人群切实体会到科技伦理的重要性，负责任地研发、使用通用模型。另一方面，需要研发更系统的解决方案和上下游技术手段，对训练数据和模型表现进行伦理调节，改善模型在各个环节的道德敏感度。

（3）要充分探索“价值敏感设计”“负责任创新”等伦理方法在通用模型语境下的可行性。不仅要将伦理规范嵌入到技术研发中去，更要考虑如何通过对模型的伦理规制与技术对人的规制的双向互动，真正探索出合乎伦理的模型的研发之路，前瞻性地规避价值缺失、价值冲突、价值违背等价值问题，保证其在合规、合伦理的框架下促进人工智能的发展。

致谢 感谢乔宇和王昊奋对本文提出的建议，感谢傅则成的前期资料收集工作。

参考文献

1 Knight J. Computer Modeling. (2017-06-11)[2022-03-06]. https://www.encyclopedia.com/social-sciences-and-law/law/crime-and-law-enforcement/computer-modeling.

2 Bommasani R, Hudson D A, Adeli E, et al. On the Opportunities and Risks of Foundation Models. San Francisco: Stanford University, 2021.

3 黄铁军, 文继荣, 刘知远, 等. 超大规模智能模型产业发展报告. 北京: 北京智源人工智能研究院, 2021.
Huang T J, Wen J L, Liu Z Y, et al. Super Large Scale Intelligent Model Industry Development Report. Beijing: Beijing Academy of Artificial Intelligence, 2021. (in Chinese)

4 徐英瑾, 陈萌. 人工智能如何“说人话”?——对于自然语言处理研究的哲学反思. 自然辩证法通讯, 2022, 44(1): 10-19.
Xu Y J, Chen M. How to make artificial intelligence capable of speaking human language? Some philosophical remarks on natural language processing. Journal of Dialectics of Nature, 2022, 44(1): 10-19. (in Chinese)

5 Clark E, August T, Serrano S, et al. All that’s ‘human’ is not gold: Evaluating human evaluation of generated text//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 7282-7296.

6 Shao J, Chen S, Li Y, et al. Intern: A new learning paradigm towards general vision. arXiv preprints, 2021: 2111.08687.

7 Goldstein J. Emergence as a construct: History and issues. Emergence, 1999, 1(1): 49-72.

8 赵斌. 充分理解涌现性, 慎重对待转基因. 科学家, 2013, 12(2): 88-89.
Zhao B. Comprehending emergence to think differently about genetically modified technology. Scientist, 2013, 12(2): 88-89. (in Chinese)

9 Samuel A L. Some studies in machine learning using the game of checkers. II—Recent progress. IBM Journal of Research and Development, 1967, 11(6): 601-617.

10 Steinhardt J. On the risks of emergent behavior in foundation models. (2021-10-18)[2022-03-05]. https://crfm.stanford.edu/commentary/2021/10/18/steinhardt.html.

11 Hendrycks D, Carlini N, Schulman J, et al. Unsolved problems in ML safety. arXiv preprints, 2021: 2109.13916.

12 Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv preprints, 2014: 1312.6199.

13 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXic preprints, 2015: 1412.6572v3.

14 Goh G, Cammarata N, Voss C, et al. Multimodal neurons in artificial neural networks. (2021-03-04) [2022-03-05]. https://distill.pub/2021/multimodal-neurons/.

15 Schwartz R, Vassilev A, Greene K, et al. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. Gaithersburg: National Institute of Standards and Technology, 2022.

16 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. arXiv preprints, 2021: 2103.00020.

17 王国豫, 梅宏. 构建数字化世界的伦理秩序. 中国科学院院刊, 2021, 36(11): 1278-1287.

Wang G Y, Mei H. Constructing ethical order of digital world. Bulletin of Chinese Academy of Sciences, 2021, 36(11): 1278-1287. (in Chinese)

18 Weidinger L, Mellor J, Rauh M, et al. Ethical and social risks of harm from language models. arXiv preprints, 2021: 2112.04359.

19 Sensity. The state of deepfakes: Landscape, threats, and impact. (2019-11-29) [2022-03-05]. https://medium.com/sensity/mapping-the-deepfake-landscape-27cb809e98bc

20 Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 3645-3650.

21 Patterson D, Gonzalez J, Le Q, et al. Carbon emissions and large neural network training. arXiv preprints, 2021: 2104.10350.

22 Bender E M, Gebru T, Mcmillan-Major A, et al. On the dangers of stochastic parrots: Can language models be too big?// Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: Association for Computing Machinery, 2021: 610-623.

23 Hey T, Tansley S, Tolle K. Jim Gray on eScience: A transformed scientific method// The Fourth Paradigm: Data-Intensive Scientific Discovery. Remond: Microsoft Research, 2009.

24 C.胡必希, 王国豫. 技术评估的方法与价值冲突. 自然辩证法研究, 2005, 21(12): 40-43.

Christoph H B, Wang G Y. The method of technology assessment and value conflict. Studies in Dialectics of Nature, 2005, 21(12): 40-43. (in Chinese)

25 Ihde D. Technology and the Lifeworld: From Garden to Earth. Bloomington: Indiana University Press, 1990.

26 Welbl J, Glaese A, Uesato J, et al. Challenges in detoxifying language models. arXiv preprints, 2021: 2109.07445.

27 Solaiman I, Dennison C. Process for adapting language models to society (PALMS) with values-targeted datasets. arXiv preprints, 2021: 2106.10328.

28 Thoppilan R, Freitas D D, Hall J, et al. LaMDA: Language models for dialog applications. arXiv preprints, 2022: 2201.08239.

29 Solaiman I, Brundage M, Clark J, et al. Release strategies and the social impacts of language models. arXiv preprints, 2019: 1908.09203.

Ethics and Governance of General Models: Challenges and Countermeasures

TENG Yan[1] WANG Guoyu[2,3*] WANG Yingchun[1]

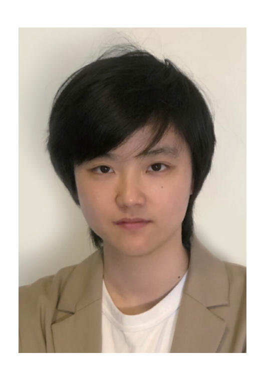
（1 Governance Research Center, Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China;

2 School of Philosophy, Fudan University, Shanghai 200433, China;

3 Center for Biomedical Ethics, Fudan University, Shanghai 200433, China）

Abstract In recent years, the general model is one of the most important development trends of artificial intelligence. With the rapidly increasing research and deployment of general models, the social and ethical effects of general models have received extensive attention. Grounded in the characteristics of general models, this article analyzes the potential ethical challenges of the models at three levels: algorithm, data, and computing power. The detailed challenges include uncertainty, truthfulness, reliability, bias, toxicity, fairness, privacy, and environmental issues. Also, through the lens of philosophy of technology, it elaborates the important reasons for the ethical challenges: the "mirroring" effect and transparency problem caused by the data-driven general models' mediation between human and the world. This relation can be depicted as "human-model (data)-world". Finally, from the perspectives of governance tools and governance mechanisms, this article reviews the current countermeasures and reflects on their limitations. It is recommended to establish an open, full-process, value-embedded ethical restraint mechanism to ensure that the general model develops in accordance with legal and ethical requirements.

Keywords artificial intelligence, general model, risk, governance, ethics of technology



滕 妍 上海人工智能实验室治理研究中心青年研究员，荷兰代尔夫特理工大学博士。主要从事科技伦理和人工智能伦理治理研究。E-mail: tengyan@pjlab.org.cn

TENG Yan Young Researcher of Governance Research Center, Shanghai Artificial Intelligence Laboratory. Ph.D. of Delft University of Technology, the Netherlands. Her main research fields cover ethics of technology, ethics and governance of artificial intelligence. E-mail: tengyan@pjlab.org.cn



王国豫 复旦大学哲学学院教授、博士生导师，复旦大学生命医学伦理研究中心、应用伦理学研究中心主任。中国自然辩证法研究会科技与工程伦理专业委员会副理事长，上海市自然辩证法研究会理事长，上海国家新一代人工智能创新发展试验区专家咨询委员会委员，中国计算机学会职业伦理与学术道德委员会共同主席。主要从事应用伦理学和科技伦理、科技哲学的研究。
E-mail: wguoyu@fudan.edu.cn

WANG Guoyu Professor, Ph.D. Supervisor of School of Philosophy, Fudan University. Director of Center for Biomedical Ethics, and also Director of Applied Ethics Center, Fudan University. Vice Chairman of Science, Technology and Engineering Ethics Committee, Chinese Society for Dialectics of Nature/Philosophy of Nature, Science and Technology (CSDN/PNST). Chairwoman of Shanghai Society of Dialectics of Nature, Member of the Expert Advisory Committee of Shanghai National Pilot Area for New Generation AI Innovation and Development, and Chairwoman of the Committee for Professional and Academic Ethics of China Computer Federation (CCF). Her main research fields cover applied ethics, ethics and philosophy of technology. E-mail: wguoyu@fudan.edu.cn

■责任编辑：张帆

*Corresponding author